biblio.ugent.be

The UGent Institutional Repository is the electronic archiving and dissemination platform for all UGent research publications. Ghent University has implemented a mandate stipulating that all academic publications of UGent researchers should be deposited and archived in this repository. Except for items where current copyright restrictions apply, these papers are available in Open Access.

This item is the archived peer-reviewed author-version of:

Machine-interpretable dataset and service descriptions for heterogeneous data access and retrieval

Anastasia Dimou, Ruben Verborgh, Miel Vander Sande, Erik Mannens, and Rik Van de Walle

In: Proceedings of the 11th International Conference on Semantic Systems, 9367, 145-152, 2015.

http://dl.acm.org/citation.cfm?id=2814873

To refer to or to cite this work, please use the citation to the published version:

Dimou, A., Verborgh, R., Vander Sande, M., Mannens, E., and Van de Walle, R. (2015). Machine-interpretable dataset and service descriptions for heterogeneous data access and retrieval. ***Proceedings of the 11th International Conference on Semantic Systems 9367*** **145-152. 10.1145/2814864.2814873**

Machine-Processable Dataset and Service Descriptions for Heterogeneous Data Access and Retrieval

Anastasia Dimou
anastasia.dimou@ugent.be

Ruben Verborgh
ruben.verborgh@ugent.be

Miel Vander Sande
miel.vandersande@ugent.be

Erik Mannens
erik.mannens@ugent.be

Rik Van de Walle
rik.vandewalle@ugent.be

Ghent University – iMinds – Multimedia Lab
Ghent, Belgium

ABSTRACT

The RDF data model allows the description of domain-level knowledge that is understandable by both humans and machines. RDF data can be derived from different source formats and diverse access points, ranging from databases or files in CSV format to data retrieved from Web APIs in JSON, Web Services in XML or any other speciality formats. To this end, vocabularies such as RML were introduced to uniformly define how data in multiple heterogeneous sources is mapped to the RDF data model, independently of their original format. This approach results in mapping definitions that are machine-processable and interoperable. However, the way in which this data is accessed and retrieved still remains hard-coded, as corresponding descriptions are often not available or not taken into account. In this paper, we introduce an approach that takes advantage of widely-accepted vocabularies, originally used to advertise services or datasets, such as Hydra or DCAT, to define how to access Web-based or other data sources. Consequently, the generation of RDF representations is facilitated, as the description of the interaction models with the original data remains independent, interoperable and granular.

Keywords

Linked Data Mapping, Data Access, Data Retrieval, RML

1. INTRODUCTION

Describing domain-level knowledge, understandable both by humans and machines, can be achieved by representing data using the RDF data model. Although, if the data is originally in other formats, its representation in RDF syntax should be obtained. Such data can originally (i) reside on *diverse, distributed locations*, (ii) be approached using *different access interfaces* and (iii) have *heterogeneous structures and formats*. In more details:

Diverse, distributed locations
: Data can reside locally, e.g., in files or in a database at the local network, or can be published on the Web.

Different access interfaces
: Data can be approached using diverse interfaces. For instance, it can be as straightforward to access the data as raw files for example. There might be metadata that describe how to access the data, as in the case of data catalogues. But it might also be required to have a dedicated access interface to retrieve the data from a repository. For instance database connectivity for databases, or different interfaces from the Web, such as Web APIs.

Heterogeneous structures and formats
: Data can be stored and/or retrieved in different structures and formats. For instance, data can originally have a *tabular structure*, (e.g., databases or CSV files), be *tree-structured* (e.g., XML or JSON format), or be *semi-structured* (e.g., in HTML).

Coping with the ever-increasing amount of data, in respect to incorporating data from multiple sources of different data formats into a common knowledge domain, is challenging but still remains complicated, despite the significant number of existing tools. To be more precise, most of the tools that generate RDF representations of some data, deploy mappings from a certain source format to RDF and from a given input. Only few provide mappings from different source formats to RDF, and even less provide independent, interoperable and machine-processable mapping definitions.

Even though the barrier of uniformly defining how to map heterogeneous data to the RDF data model has been addressed [11], even more generic application still can not be built because data access and retrieval remains hard-coded. To be more specific, uniform, machine-processable mapping definitions indicate how triples should be generated in a generic way for all possible different input sources. Those mapping definitions contain references to an input data source, which are case-specific and, thus, defined using formulations relevant to the corresponding data format, e.g., XPath for data in XML format. However, as the data retrieval remains out of the mapping definitions' scope, it ends up being hard-coded in the corresponding implementations. While this is not a major problem when local, custom, or input-specific data is considered, the situation aggravates

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

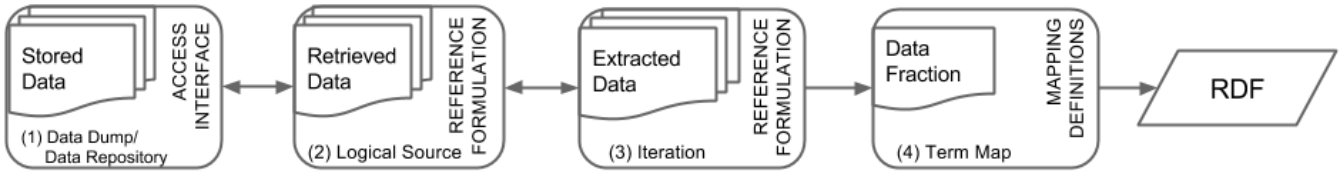

Figure 1: Data retrieval and mapping to the RDF data model.
A triple consists of RDF Terms which are generated by Term Maps (4). Those Term Maps are defined with a *mapping definition vocabulary* and are instantiated with data fractions referred to using a *reference formulation* relevant to the corresponding data format. Those fractions are derived from data extracted at a certain iteration (3) from a logical source (2). Such a logical source is formed by data retrieved from a repository (1) which is accessed as defined using the corresponding *dataset or service description vocabulary*.

when data from multiple heterogeneous data sources, accessed via different interfaces, is required to be retrieved and mapped to the RDF data model.

Vocabularies which are originally used to advertise datasets or services (e.g., DCAT[1] or Hydra[2]) and to enable applications to easily consume the underlying data, exist. These same vocabularies can be used to specify how to access and, subsequently, retrieve data sources, available on the Web or not and generate their RDF representation. This way, the description that captures how to access the data becomes machine-processable, as the mapping descriptions are, enabling even more generic implementations. However, access descriptions with such vocabularies are not taken into account and are not aligned with the vocabularies used to describe the mapping definitions.

In this paper, we introduce an approach that exploits W3C-recommended or widely-accepted vocabularies originally used to advertise datasets or services, e.g., DCAT or SPARQL-SD[3], to define how to access data sources, available on the Web or not, and generate their RDF representation. Our contribution is twofold: (i) on the one hand, we review different vocabularies of interfaces that describe how to access data and we define how data sources can be instantiated using those descriptions; (ii) on the other hand, we define how such access interface descriptions can be aligned with a mapping language and we extend the generic mapping language RML[4], to properly handle such input sources.

The remainder of the paper is structured as follows: Section 2 elucidates the retrieval steps required to obtain the data whose semantic representation is desired. Section 3 reviews related works. Section 4 describes machine processable service and dataset descriptions for different cases and Section 5 provides details regarding how RML was extended to take them into consideration. Finally, Section 6 concludes with the outcomes of this work.

2. ACCESS, RETRIEVE AND MAP DATA TO THE RDF DATA MODEL

In order to describe domain-level knowledge, understandable both by humans and machines the RDF data model can be considered. Although, if the data is originally in other formats, their representation in RDF syntax should be obtained. In this section, we explain the retrieval and extraction steps required to obtain the data whose semantic representation is desired. Figure 1 illustrates how data is accessed and retrieved from their original repositories and how further data fractions are extracted to finally obtain the desired RDF dataset.

Data is stored in different repositories residing sometimes in different locations. Those repositories can be found e.g., locally, on a network, or on the Web. For instance, data can be available as raw files, databases or Web resources, or files listed in catalogues[5]. To retrieve data from a repository, an *access interface* is required (Step 1) to handle the interaction. Data can be approached using diverse interfaces. For instance, database connectivity, such as Open DataBase Connectivity (ODBC) to access data residing in a database. But data on the Web can also be retrieved using different interfaces, such as Web APIs[6] or Web services.

Once the *retrieved data* is obtained (Step 2), from one or more repositories, one or more *Logical Sources* are formed. Such a *Logical Source* contains data in a certain structure and format, e.g., CSV, XML or JSON. This data source is what mapping languages, such as RML, consider for the mapping definitions. How this data source is retrieved is out of scope for vocabularies focused on specifying the mapping definitions. If the original repository is a raw file, the *Logical Source* may coincide. Further data fragmentation and extraction requires references relevant to the data format (i.e., its corresponding *Reference Formulation*).

As *mapping definitions* are meant to be applied recursively to data fragments extracted from the *Logical Source*, an *iterator* is required. The iteration pattern is also defined in a formulation relevant to the *Logical Source*. The *iterator* runs over the *Logical Source*, extracting data fragments (Step 3). For instance, an *iterator* running over a CSV file extracts a row of the CSV at each iteration. In case the iteration pattern applies to the complete *Logical Source*, the *Iteration* fragment coincides with the *Logical Source*.

For each *Iteration* further data fragmentation occurs (Step 4) to extract the exact *Data fraction(s)* used to instantiate a *Term Map* which, in its turn, generates the corresponding RDF term. For the aforementioned CSV example, such a data fraction is the value of a column from a given row extracted at a certain *Iteration*. At the end, the corresponding RDF representation of the *Logical Source* is obtained (Step 5).

[1] www.w3.org/TR/vocab-dcat/
[2] http://www.w3.org/ns/hydra/spec/latest/core/
[3] http://www.w3.org/TR/sparql11-service-description/
[4] http://rml.io
[5] e.g.,http://open-data.europa.eu/en/data/dataset/cordisfp7projects
[6] e.g.,https://biblio.ugent.be/publication/{id}?format={format}

3. RELATED WORK

To the best of our knowledge, there is no mapping solution that takes into consideration diverse dataset and services descriptions to access the data described. Most existing solutions consider data derived from a certain source format and from a given input, which, in most cases, is a local file.

3.1 Mapping Languages

For relational databases, different mapping languages are defined [13]. Indicatively mentioned, the Triplify [2] which is based on mapping HTTP-URI requests onto relational database queries, and the Sparqlification Mapping Language (SML) [24] which declaratively defines mappings based on SQL views and SPARQL construct queries, do not specify how the input data source is retrieved from the corresponding database within the mapping definitions. Among those language, only D2RQ [7], which is described in more details in Section 4.4, defines how the database connectivity should be specified. To the contrary, the W3C recommended R2RML [8], does not provide any database connectivity descriptions, as it considers such description out of the vocabulary scope.

Mapping languages were also defined to support conversion from data in CSV and spreadsheets to the RDF data model. XLWrap's mapping language [17] converts data in various spreadsheets to RDF. XLWrap's mapping language does not describe alternative access descriptions, as a file is always the expected data source. However, which file exactly is expected, is specified within the mapping definition as follows `[ ] xl:fileName ''files/example.xls''`. Similarly, TARQL that follows a query-based approach, considers CSV files as input. Such a file is also defined within the query that acts as mapping definition. In TARQL language [6], the mapping definitions have SPARQL syntax, thus the input CSV file is defined as follows `SELECT ... FROM <file:example.csv>`. Last, the declarative OWL-centric mapping language Mapping Master's M^2 [23] which converts data from spreadsheets into OWL, does not specify at all within the mapping definitions the input source.

A larger variety of solutions exist to map data in XML format to RDF, but tools mostly rely on existing XML solutions, such as XSLT (e.g., Krextor[7] and AstroGrid-D[8]), XPath (e.g., Tripliser[9]), and XQuery (e.g., XSPARQL[10]). None of them though defines neither how the input source should be specified, nor has RDF syntax which would allow them to be combined with dataset and service access descriptions.

Last, among the tools that provide mappings from different source formats to the RDF data model, e.g., Datalift[11], OpenRefine[12], RDFizers[13] or Virtuoso Sponger[14], none relies on independent generic mapping definitions. Instead those tools employ separate source-centric approaches for each of the formats they support which are hard-coded in the corresponding implementation. The only generic language that exists and allows any type of input source is RML [11], which is described in more details in Section 5.1.

[7] https://trac.kwarc.info/krextor/

[8] http://www.gac-grid.de/project-products/Software/XML2RDF.html

[9] http://daverog.github.io/tripliser/

[10] http://www.w3.org/Submission/xsparql-language-specification/

[11] http://datalift.org/

[12] http://openrefine.org/

[13] http://simile.mit.edu/wiki/RDFizers

[14] http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtSponger

3.2 Dataset and Service Descriptions

Different dataset and service descriptions exist, which describe how to access data. Rather than reinventing such descriptions for the purpose of data access, we aim to reuse existing work, which we summarize and discuss in the following paragraphs. Dataset descriptions could refer to data catalogues to Lined Data sets, or to specific type of data, e.g., tabular data. In the former case, the W3C recommended vocabulary, DCAT [20], is defined which is more thoroughly described in Section 4.1. In the later case, the VOID vocabulary [1] is considered which defines how to describe metadata for RDF datasets to improve their discoverability. Among the metadata which can be specified with the VOID vocabulary, it is also *access metadata*. The VOID vocabulary allows to specify as access interface (i) SPARQL endpoints, (ii) RDF data dumps, (ii) root resources, (iv) URI lookup endpoints and (v) OpenSearch description documents. In the same context, the CSV on the Web Working Group[15] aims to define a case-specific metadata vocabulary for Tabular data on the Web [25] which, at its current state, only allows data dumps as access interface.

As far as service descriptions is concerned, and in respect to accessing data in RDF syntax, besides the VOID vocabulary, there is the W3C recommended SPARQL-SD [26], which is described in more details in Section 4.3. Regarding database connectivity, there are no dedicated vocabularies. Descriptions in the frame of mapping languages, e.g., D2RQ, which is also described in more details in Section 4.4, prevail. However, regarding Web APIs and Services, different vocabularies were defined, thus we review below the state of the art.

The Web Service Description Language (WSDL) [5] describes the possible interactions, messages and the abstract functionality provided by Web services. [14] describes its representation in RDF and in OWL, as well as a mapping procedure for transforming WSDL descriptions into RDF. Semantic Annotations for WSDL (SAWSDL) [16] is one of the first attempts to offer semantic annotations for Web services. Later on, an adaptation for generic HTTP interfaces was proposed [21].

The OWL for Services (OWL-S) [22], the Web Service Modeling Ontology (WSMO) [9] and the WSMO-Lite [27] are alternative ontologies, defined for modelling Web services. The OWL-S ontology also focuses on input and output parameters, as SAWSDL. The WSMO ontology is an alternative to OWL-S, although there are substantial differences between the two approaches [19]. The WSMO ontology employs a single family of layered logic languages [10]. However, when expressed in RDF syntax, WSMO expressions become similarly unintegrated and hence not self-descriptive as OWL-S expressions. The WSMO-Lite ontology extends SAWSDL with conditions and effects. hRESTS [15] uses microformats to add machine-processable information to human-readable documentation, while its ontology[16] extends the WSMO-Lite ontology[17]. Last, MicroWSMO extends hRESTS and adopts the WSMO-Lite service ontology for expressing concrete semantics. For our purposes, we mostly need the interface description part of the above possibilities, since our goal is access to the services rather than, for instance, composition.

[15] http://www.w3.org/2013/csvw/wiki/Main_Page

[16] http://www.wsmo.org/ns/hrests/

[17] http://www.wsmo.org/ns/wsmo-lite/

4. DESCRIBING INTERFACES TO ACCESS HETEROGENEOUS DATA SOURCES

Even though the barrier of uniformly mapping heterogeneous data to the RDF data model has been overcome, with uniform, machine-processable mapping definitions and case specific references to the input data source, depending on its format, data access and retrieval remains hard-coded in the implementation. Data can reside locally or on the Web. Accessing the data might be straightforward, as in the case of files locally stored. However, in most cases, dedicated interfaces are required. Such dataset and service descriptions can be: (i) dataset's metadata descriptions, (ii) Hypermedia-driven Web APIs or Web services, (iii) SPARQL services, (iv) database connectivity.

In the previous section (Section 3), we review vocabularies describing interfaces for accessing datasets and services that enable agents to retrieve the underlying data. For each type of interface, a corresponding W3C recommended is described in more details below. In case there is no such vocabulary, a widely-used one is mentioned. The list is not exhaustive, it rather has an indicative exemplary purpose, aiming to capture the most common cases. Any of the dataset or service descriptions could be replaced by other corresponding ones and new can be brought into consideration.

4.1 Metadata describing the access interface

Data can be published either independently, as data dumps, or in the frame of a data catalogue. DCAT [20] is the W3C recommended vocabulary used to describe datasets in data catalogs. The DCAT vocabulary provides machine-readable metadata that enables applications to easily consume them. The DCAT vocabulary does not make any assumptions about the format of the datasets described in a catalog; format-specific information is out of scope. The DCAT namespace is `http://www.w3.org/ns/dcat#` and the preferred prefix is *dcat*.

The DCAT vocabulary defines *dcat:Catalog* that represents a dataset catalog, *dcat:Dataset* that represents a dataset in the catalog, while *dcat:Distribution* represents an accessible form of a dataset, e.g., a downloadable file, an RSS feed or a Web service that provides the data. DCAT considers as a dataset a collection of data, published or curated by a single agent, and available for access or download in one or more formats. Thus, a certain distribution is the minimum that a mapping processor requires. Directly downloadable distributions contain a *dcat:downloadURL* reference, for instance:

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .

<#DCAT_source>
  a dcat:Dataset ;
  dcat:distribution [
    a dcat:Distribution;
    dcat:downloadURL <http://example.org/file.xml>].
```

Listing 1: DCAT access metadata description

A *dcat:Distribution* might not be directly downloadable though. For instance, a dataset might be available via an API and the API, in its turn, can be defined as an instance of a *dcat:Distribution*. In this case, it is recommended to use *dcat:accessURL* instead of *dcat:downloadURL*. However, access-specific properties, e.g., for API descriptions, is not defined by DCAT itself. Thus a client does not know how to interact with the mentioned interface, the API in this case. Due to DCAT shortcoming to entirely describe indirectly accessed Web sources, other vocabularies focused on describing specific interfaces could be considered instead.

4.2 Hypermedia-Driven Web APIs

For the description of hypermedia-driven Web APIs, the Hydra Core Vocabulary [18], a lightweight vocabulary used to specify concepts commonly used in Web APIs, is published by the Hydra W3C Community Group[18]. The Hydra vocabulary provides machine-processable descriptions which enable a server to advertise valid state transitions to a client. Thus, the server is decoupled from the client which can use this information to construct valid HTTP requests to retrieve the data. The Hydra namespace is `http://www.w3.org/ns/hydra/core#` and the preferred prefix *hydra*.

An instance of the *hydra:ApiDocumentation* class describes a Web API, by providing its title, short description, main entry point and additional information about status codes that might be returned. The Hydra vocabulary enables the API's main entry point to be discovered automatically, when it is not known or specified, if the API publisher marks his responses with a special HTTP link header. A client looks for a link header with a relation type *hydra:apiDocumentation* and, this way, obtains a *hydra:ApiDocumentation* defining the API's main entry point.

The *dcat:accessUrl* of a *dcat:Distribution* instance can point to a resource described with the Hydra vocabulary, informing potential agents how valid HTTP requests should be performed. The Hydra vocabulary can be used both to describe (i) static IRIs, and (ii) dynamically generated IRIs, e.g., Listing 2. A template valued IRI containing variables, is described as a *hydra:IriTemplateMapping* instance whose values depend on information only known by the client.

```
@prefix hydra : <http://www.w3.org/ns/hydra/core#> .

<#API_source>
  a hydra:IriTemplate
  hydra:template
    "https://api.twitter.com/1.1/followers/ids.json?
     screen_name={name}";
  hydra:mapping [
    a hydra:TemplateMapping ;
    hydra:variable "name";
    hydra:required true ] .
```

Listing 2: Template-valued Web API description

Web APIs often split a collection of data into multiple pages. In Hydra, this is described with an instance of the *hydra:PagedCollection* that contains information regarding the total number of items, the number of items per page and the first, the next and the last page. Such a *hydra:PagedCollection* instance follows:

```
@prefix hydra : <http://www.w3.org/ns/hydra/core#> .

<#API_source>
  a hydra:PagedCollection ;
  hydra:apiDocumentation <#HydraDocumentation> ;
  hydra:itemsPerPage "100" ;
  hydra:firstPage "/comments?page=1";
  hydra:lastPage "/comments?page=10" .
```

Listing 3: Hydra Paged Collection description

[18] `http://www.w3.org/community/hydra/`

Web Services

Web services played an important part in the initial Semantic Web vision [3]. However, Web Services were surpassed in popularity by Web APIs and at the moment, most of the Web-based solutions prefer the later. Thus, a detailed example for Web Service descriptions is not provide, even though such a description could equally be considered.

4.3 SPARQL services

For the description of SPARQL endpoints, W3C recommends the SPARQL Service Description vocabulary (SPARQL-SD) [26]. SPARQL-SD provides a list of features of a SPARQL service and their descriptions, made available via the SPARQL 1.1 Protocol for RDF. The SPARQL-SD namespace is `http://www.w3.org/ns/sparql-service-description#` and the preferred prefix is *sd*.

An instance of *sd:Service* represents a SPARQL service made available via the SPARQL protocol. A *sd:Service* refers to a default dataset, described as an instance of the *sd:Dataset* that represents an RDF dataset comprised of a default graph (an instance of *sd:Graph*) and zero or more named graphs (an instance of *sd:NamedGraph*). A collection of graphs is described as instances of *sd:GraphCollection*. Last, SPARQL-SD defines *sd:Language* whose instances represent one of the SPARQL languages (e.g., *sd:SPARQL11Query*).

```
<#SPARQL_source>
  a sd:Service ;
  sd:endpoint <http://dbpedia.org/sparql/> ;
  sd:supportedLanguage sd:SPARQL11Query ;
  sd:resultFormat
    <http://www.w3.org/ns/formats/SPARQL_Results_XML>.
```

Listing 4: SPARQL Service Description

Similarly to Hydra, a *sd:Service* instance could be used to clarify *dcat:accessUrl*, allowing potential agents to know how to perform the corresponding HTTP requests.

4.4 Database Connectivity

For the description of database connectivity, corresponding descriptions from the D2RQ mapping language can be considered [7]. D2RQ is a declarative mapping language for describing the relation between a relational database schema and RDFS vocabularies or OWL ontologies. The D2RQ namespace is `http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#` and the preferred prefix *d2rq*.

A *d2rq:Database* instance defines a JDBC connection to a local or remote relational database. Instances of d2rq:Database, annotated with its properties to specify the JDBC connection properties, can be used to describe the access to a database. An instance of such database description has as follows:

```
@PREFIX d2rq:
  <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#>

<#DB_source>
  a d2rq:Database;
  d2rq:jdbcDSN "jdbc:mysql://localhost/example";
  d2rq:jdbcDriver "com.mysql.jdbc.Driver";
  d2rq:username "user";
  d2rq:password "password" .
```

Listing 5: D2RQ database connectivity description

The D2RQ database connectivity description is focused on relational databases and serves the needs of an exemplary case of this work. Corresponding vocabularies for other type of databases (e.g., NoSQL) can be taken into consideration.

5. ALIGNING DATA ACCESSING AND MAPPING TO RDF DESCRIPTIONS

In this section, we define in details how heterogeneous dataset and service descriptions can be taken into consideration to access data and instantiate RML Logical Sources. Those Logical Sources contain data whose representation in RDF syntax is desired. For our use case, we align the aforementioned descriptions with RML mapping language. In Section 5.1, we introduce the language, and in Section 5.2, we concretely define how RML Logical Sources are obtained via such dataset and service descriptions in the frame of RML. Finally, in Section 5.3 we introduce required extension to the RML mapping language to entirely support such logical sources. Detailed documentation regarding different access interfaces supported by RML is available at `http://rml.io/RML_Input.html`

5.1 RML

RML [11] extends R2RML [8], the W3C-recommended mapping language for defining mappings of data in relational databases to the RDF data model, by broadens its scope. RML covers also mappings from data sources in different (semi-)structured formats, such as CSV, XML, and JSON.

```
1  <#Person> rml:logicalSource <#InputX> ;
2    rr:subjectMap [
3     rr:template "http://ex.com/{ID}";
4     rr:class foaf:Person ];
5    rr:predicateObjectMap [
6     rr:predicate foaf:account;
7     rr:objectMap [ rr:parentTriplesMap <#TwitterAcount> ] ] .
8
9  <#TwitterAcount> rml:logicalSource <#InputY> ;
10   rr:subjectMap [
11     rr:template "http://ex.com/{account_ID}";
12     rr:class foaf:OnlineAccount ];
13   rr:predicateObjectMap [
14    [ rr:predicate foaf:accountName;
15      rr:objectMap [ rml:reference "name"] ],
16    [ rr:predicate foaf:accountServiceHomepage;
17      rr:objectMap [ rml:reference "resource"] ] .
```

Listing 6: RML mapping definitions

RML documents contain rules defining how the input data is represented in RDF. An RML document (e.g., Listing 6) contains one or more Triples Maps (line 1 and 9). A Triples Map defines how triples of the form (subject, predicate, object) are generated and consists of three main parts: the Logical Source, the Subject Map and zero or more Predicate-Object Maps. The Subject Map (line 2, 10) defines how unique identifiers (URIs) are generated for the resources and is used as the subject of all RDF triples generated from this Triples Map. A Predicate-Object Map (line 5 and 13) consists of Predicate Maps, which define the rule that generates the triple's predicate (line 6, 14 and 16) and Object Maps (line 14 and 16) or Referencing Object Maps (line 6), which define how the triple's object is generated.

```
<#InputX>
  rml:source "/.../.../file.csv" ;
  rml:referenceFormulation ql:CSV.
```

Listing 7: RML Logical Source definition - local file

A Logical Source (e.g., Listing 7) is used to determine the input source (line 2) with the data to be mapped and how to refer to them (line 3). RML deals with different data serialisations which use different ways to refer to their elements/ob-

jects. RML considers that any reference to the Logical Source should be defined in a form relevant to the input data, e.g. XPath for XML files or JSONpath for JSON files. To this end, the Reference Formulation (line 3) declaration is stated indicating the formulation (for instance, a standard or a query language) used to refer to its data. At the current version of RML, the `ql:CSV`, `ql:XPath`, `ql:JSONPath` and `ql:CSS3` Reference Formulations are predefined, but not limited.

5.2 Dataset and service access descriptions as RML Logical Sources

RML provides a generic way to define the mappings that is easily transferable to cover references to other data structures. RML needs to deal with different data serialisations which use different ways to refer to their data fragments. Since RML aims is generic, there is no uniform way of referring to these data fragments. RML considers that any reference to the source should be defined in a form relevant to the input data, e.g. XPath for XML files or JSONpath for JSON files. This is defined using the Reference Formulation (rml:referenceFormulation) declaration that indicates the formulation (for instance, a standard or a query language) used to refer to source's data fragments.

However, the RML specification is focused on the rules defining how to generate the RDF data. RML considers a given original data input but the way this input is retrieved remains out of scope, in the same way as it remains out of scope for R2RML specification how the SQL connection is established. The input data is specified with the Logical Source, as well as how to refer to this data, but not how to access and retrieve this data. Namely, the Logical Source consists of some data without further defining how to retrieve the data.

The access descriptions, that advertise services or datasets, could be considered as the Triples Map's Source (*rml:source*). For instance, the Logical Source specified at Listing 6 for the `<#Person>` Triples Map, instead of having been specified as a local file (Listing 7), it could have been published on a data catalogue and, thus, it is an instance of *dcat:Distribution*. The corresponding description then would be as follows:

```
1  <#InputX>
2    rml:source [ a dcat:Distribution ;
3      dcat:downloadURL "http://ex.com/file.csv" ];
4    rml:referenceFormulation ql:CSV .
```

Listing 8: Data dump in catoague as Input Source

For the other Triples Map, `<#TwitterAccount>`, the data to be mapped might be derived from a user's twitter account, and could have been stored locally in a file retrieved at some point from the Twitter API, or the request could have been hard-coded in the implementation. Nevertheless the required request could have just been described using the Hydra vocabulary or could have been provided using directly the resource advertising the API. In the aforementioned example of Listing 6, the Logical Source for the `<#TwitterAccount>` Triples Map could have been described as follows:

```
1  <#InputY> a hydra:IriTemplate
2    hydra:template "https://api.twitter.com/1.1/followers/ids.
         json?screen_name={name}";
3    hydra:mapping [
4      hydra:variable "name";
5      hydra:required true ] .
```

Listing 9: Web API as Input Source

5.3 RML Referencing Object Map and Heterogeneous Data Retrieval

The use of data derived from such a *Logical Source*, formed by instantiating an access description, is straightforward in most cases. Dataset and service descriptions either are derived from *data owners/publishers* or explicitly defined by *data publishers/consumers*. Mapping processors take them into consideration to be informed regarding how to access the data and instantiate the *Logical Sources*. The access description might be static or dynamic. If dynamically created, it is often required to instantiate a template, e.g., a URI template or a SQL/SPARQL query template. The values to instantiate the template might be provided by the user who executes the mapping or the variables might be replaced with values derived from another input source, as it occurs in the case of *Referencing Object Maps*.

Binding condition

A Referencing Object Map (line 3) allows using the subject of another Triples Map(line 9) as the objects generated by a Predicate Object Map. and the two Triples Maps may be based on different Logical Sources.

A Referencing Object Map might have a *template-valued* input source that requires one or more values to fill in the template. In order to address this issue, we introduced the Binding Condition. The Binding Condition specifies how the Logical Source of the Referencing Object Map is instantiated either with a value retrieved from the input source that is currently mapped, or with a constant value. In the first case, a reference that exists in the Logical Source of the Triples Map that contains the Referencing Object Map is provided. In the later case, a constant value is provided. If the Referencing Object Map's Logical Source has more than one variables required, equal number of Binding Conditions is expected.

```
1  <#Person>
2    rr:predicateObjectMap [ rr:predicate foaf:account;
3      rr:objectMap [
4        rr:parentTriplesMap <#TwitterAcount> ;
5        crml:bindCondition [
6          rml:reference "id" ;
7          crml:condition "name" ]]].
8
9  <#InputY> rml:source [
10   a hydra:IriTemplate
11   hydra:template "https://api.twitter.com/1.1/followers/ids.
         json?screen_name={name}";
12   hydra:mapping [
13     hydra:variable "name";
14     hydra:required true ] ];
15   rml:referenceFormulation ql:JSONPath.
```

Listing 10: RML Binding Condition

Detailed documentation regarding the *RML Bind Condition* is available at `http://rml.io/cRML_bindCondition.html`.

5.4 Implementation

An RMLProcessor can be implemented using two alternative models: (i) *mapping-driven*, where the processing is driven by the mapping module; or (ii) *data-driven*, where the processing is driven by the extraction module [12]. When the RML mappings are processed, the mapping module deals with the mapping definitions execution, while the extraction module deals with the target language expressions (expression using the corresponding *Reference Formulation*).

On the *mapping-driven* occasion, the *mapping module* re-

quests an extract of data from the *extraction module*, considering the iteration pattern specified at the *Logical Source*. On the *extraction-driven* occasion, an extract of data is passed to the mapping module, which applies the applicable mapping definitions for this particular extract.

A new additional independent *retrieval module* is introduced which deals with the retrieval of data that form the *Logical Source*. The *retrieval module* relies on the access description to retrieve the data. The access description can be provided either by the *data owner/publisher* or by the *data consumer/publisher*, but in both cases the descriptions are equally treated. Moreover, if the access description is dynamically generated either user input is taken into consideration to bind the template variables with values, or values derived from another *Logical Source* Overall, none of the two aforementioned cases (*mapping-* or *data-driven*) is affected by the way the data is retrieved. A separate project, *RMLDataRetrieval*, is introduced as part of the RMLProcessor[19] that deals with data retrieval. The *RMLDataRetrieval* project is included in the RMLProcessor and currently supports most of the access interfaces described in Section 4.

5.5 Discussion

Being able to consider diverse access interfaces facilitates the description of the interaction models while the original data remains independent, interoperable and granular. In the same time, the alignment of dataset and service descriptions with the mapping definitions as proposed in this work, demands certain clarifications. Firstly, it is required to address the cases where both the *data publishers/consumer* provides access descriptions and the *data publishers/owner*. Then, it is required to elucidate how the mapping definitions should be defined depending on whether the database connectivity description is specified or not. Last, the role of RDF data, accessed e.g., via SPARQL-SD, as input source should be clarified.

Published vs. Defined Data Access Description

If the service provides access metadata, the *data publishers/consumer* can just point to the resource that describes them. If not, the minimum required information for each access interface should be defined. In case the data access is described by the *data publisher/consumer*, its description prevails over the one provided by the *data publisher/owner*. For instance, in the case of a *hydra:PagedCollection* instance, the *data publisher/consumer* might define at the data access description that a hundred items per page should be returned and five pages should be taken into consideration. If the publishing service returns an answer that contains ten pages of data, only the five of them should be mapped. If the *data publisher/consumer* does not specify the pages, all of them will be considered for mapping.

Database connectivity description with RML Logical Source and R2RML Logical Table

A Logical Source (e.g., Listing 7) extends R2RML's Logical Table and is used to determine the input source with the data to be mapped and how to refer to them. The R2RML Logical Table definition determines a database's table, using the Table Name. Nevertheless, how the connection to the database is achieved is not specified at the R2RML specification, since it remains out of its scope. Moreover, R2RML is specific for SQL databases, while a D2RQ description may refer to other databases too. In order to deal with database retrieval, we rely on such descrptions as specified in another mapping language, namely, D2RQ.

Thus, in order to take advantage of database connectivity descriptions and, thus, deal with data retrieval from databases, connectivity descriptions are considered. For instance, in the case of an SQL query against the table *DEPT* of a database, the R2RML *Logical Table* would have been defined as follows:

```
[ ] rr:logicalTable [ rr:sqlQuery """
  SELECT DEPTNO, DNAME, LOC,
  (SELECT COUNT(*) FROM EMP WHERE EMP.DEPTNO=DEPT.DEPTNO)
  AS STAFF FROM DEPT; """ ].
```

Listing 11: R2RML Logical Table

However, if a database is specified, the *Logical Table* should be expressed as an instance of its broader *Logical Source* and the corresponding database connectivity description should be provided, as follows (`<#DB_source>` was defined at Listing 5):

```
[ ] rr:logicalSource [
  rml:query """
    SELECT DEPTNO, DNAME, LOC,
    (SELECT COUNT(*) FROM EMP WHERE EMP.DEPTNO=DEPT.DEPTNO)
    AS STAFF FROM DEPT; """ ;
  rml:source <#DB_source> ].
```

Listing 12: RML Logical Source for Database Input

SPARQL service as Logical Source

Having a SPARQL-SD as *Logical Source* might seem contradictory, as the data it contains are already semantically annotated and, thus, it is not required to be mapped to the RDF data model. However, there are cases that a resource is already defined and assigned a URI and no new URI is willing to be generated, it is rather preferred to point to this resource. For instance, there is a CSV file containing some data related to addresses, and among others, there is a column with country names and a certain cell might contain e.g., Belgium. Instead of generating a new resource with a new unique URI, a *Referencing Object Map* should be defined instead, whose *Logical Source* is the result of executing a query against, for instance, DBpedia endpoint, whose access description is defined as a *sd:Service*, as follows:

```
<#Address> rr:predicateObjectMap [
  rr:predicate ex:country ;
  rr:objectMap [ rr:parentTriplesMap <#Country> ] ].

<#Country> rr:logicalSource [
  rml:query """
    SELECT distinct ?Concept
    WHERE {
      ?Concept a <http://dbpedia.org/ontology/Country> ;
        rdfs:label "Belgium"@en}  """ ;
  rml:source <#DBPedia> ] ;
  rr:subjectMap [ rml:reference "/sparql/results/result/
     binding/uri"].

<#DBPedia>
  sd:endpoint <http://dbpedia.org/sparql/> ;
  sd:supportedLanguage sd:SPARQL11Query ;
  sd:resultFormat
    <http://www.w3.org/ns/formats/SPARQL_Results_XML>.
```

Listing 13: SPARQL Endpoint for Input Source

[19] https://github.com/mmlab/rmlprocessor

6. CONCLUSIONS AND FUTURE WORK

In this paper, we introduce an approach that exploits vocabularies originally used to advertise services or datasets, to define how to access Web-based or other data sources. The alignment of access descriptions with mapping definitions provides a modular but robust way of specifying how the input data is retrieved and mapped to the RDF data model. With the proposed solution, the generation of RDF representations is facilitated, as the description of the access interface for the original data source remains independent, interoperable and granular. In the future, we will investigate the incorporation of custom third-party services.

7. ACKNOWLEDGEMENTS

Research activities described in this paper were funded by Ghent University, iMinds (Interdisciplinary research institute for Technology founded by the Flemish Government), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the EU.

8. REFERENCES

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note, Mar. 2011.

[2] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller. Triplify: Light-weight Linked Data Publication from Relational Databases. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09. ACM, 2009.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 2001.

[4] A. Brown. Web Services Glossary. W3C Working Group Note, Feb. 2004.

[5] E. Christensen, F. Curbera, G. Meredith, and S. Weerawarana. Web Services Description Language (WSDL) 1.1. W3C Note, Mar. 2001.

[6] R. Cyganiak. Tarql SPARQL for Tables: Turn CSV into RDF using SPARQL syntax. Technical report, Aug. 2015.

[7] R. Cyganiak, C. Bizer, J. Garbers, O. Maresch, and C. Becker. The D2RQ Mapping Language. Technical report, Mar. 2012.

[8] S. Das, S. Sundara, and R. Cyganiak. R2RML: RDB to RDF Mapping Language. Working group recommendation, W3C, Sept. 2012.

[9] J. de Bruijn, C. Bussler, J. Domingue, D. Fensel, M. Hepp, U. Keller, M. Kifer, B. KÃűnig-Ries, J. Kopecky, R. Lara, H. Lausen, E. Oren, A. Polleres, D. Roman, J. Scicluna, and M. Stollberg. Web Service Modeling Ontology (WSMO). W3C Member Submission, June 2005.

[10] J. de Bruijn, D. Fensel, U. Keller, M. Kifer, H. Lausen, R. Krummenacher, A. Polleres, and L. Predoiu. Web Service Modeling Language (WSML). W3C Member Submission, June 2005.

[11] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In *Workshop on Linked Data on the Web*, 2014.

[12] A. Dimou, M. Vander Sande, J. Slepicka, P. Szekely, E. Mannens, C. Knoblock, and R. Van de Walle. Mapping hierarchical sources into RDF using the RML mapping language. In *Proceedings of the 8th IEEE International Conference on Semantic Computing*, 2014.

[13] M. Hert, G. Reif, and H. C. Gall. A comparison of RDB-to-RDF mapping languages. I-Semantics '11, pages 25–32. ACM, 2011.

[14] J. Kopecký. Web Services Description Language (WSDL) Version 2.0: RDF Mapping. W3C Working Group Note, June 2007.

[15] J. Kopecký, K. Gomadam, and T. Vitvar. hrests: An html microformat for describing restful web services. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. IEEE Computer Society, 2008.

[16] J. Kopecký, T. Vitvar, C. Bournez, and J. Farrell. Sawsdl: Semantic annotations for wsdl and xml schema. *IEEE Internet Computing*, 11, 2007.

[17] A. Langegger and W. Wöß. XLWrap – Querying and Integrating Arbitrary Spreadsheets with SPARQL. In *Proceedings of 8th ISWC*, pages 359–374. Springer, 2009.

[18] M. Lanthaler. Hydra Core Vocabulary. Unofficial Draft, June 2014.

[19] R. Lara, D. Roman, A. Polleres, and D. Fensel. A Conceptual Comparison of WSMO and OWL-S. In *Web Services*, volume 3250 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004.

[20] F. Maali and J. Erickson. Data Catalog Vocabulary (DCAT). W3C Recommendation, Jan. 2014.

[21] M. Maleshkova, J. Kopecký, and C. Pedrinaci. Adapting SAWSDL for Semantic Annotations of RESTful Services. In *On the Move to Meaningful Internet Systems: OTM 2009 Workshops*, volume 5872 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2009.

[22] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, B. Parsia, T. Payne, E. Sirin, N. Srinivasan, and K. Sycara. OWL-S: Semantic Markup for Web Services. W3C Member Submission, Nov. 2004.

[23] M. J. O'Connor, C. Halaschek-Wiener, and M. A. Musen. Mapping Master: a flexible approach for mapping spreadsheets to OWL. ISWC'10, 2010.

[24] C. Stadler, J. Unbehauen, P. Westphal, M. Ahmed Sherif, and J. Lehmann. Simplified RDB2RDF Mapping. In *Workshop on Linked Data on the Web*, 2015.

[25] J. Tennison, G. Kellogg, and I. Herman. Model for Tabular Data and Metadata on the Web. W3C Working Draft, Apr. 2015.

[26] G. Todd Williams. SPARQL 1.1 Service Description. W3C Recommendation, Mar. 2013.

[27] T. Vitvar, J. Kopecký, J. Viskova, and D. Fensel. WSMO-Lite Annotations for Web Services. In *The Semantic Web: Research and Applications*, volume 5021 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2008.